
Faire *autorité* sur le web.

Analyse des stratégies de mise en visibilité et de contrôle de l'information géographique par les institutions publiques en Amérique Latine.

Matthieu Noucher¹, Pierre Gautreau²

1. UMR ADESS, CNRS – Université Bordeaux Montaigne – Université de Bordeaux
Maison des Suds, Esplanade des Antilles, 33607 Pessac – matthieu.noucher@cnrs.fr

2. UMR PRODIG, Université Paris 1 Panthéon Sorbonne
2 rue Valette – 75005 Paris – pierre.gautreau@univ-paris1.fr

RESUME. Les référentiels géographiques institutionnels sont aujourd'hui concurrencés par une offre en données qui ne cesse de se démultiplier sur Internet. Face à de nouveaux acteurs (des multinationales comme Google aux communautés de pratique comme OpenStreetMap), les producteurs historiques doivent réaffirmer la légitimité de leurs données. Cette proposition vise à analyser leurs stratégies pour se (re)positionner sur le web. A partir d'un corpus de 1674 sites diffusant des données environnementales en libre-accès au Brésil, en Bolivie et en Argentine, nous analysons le cas particulier des données géographiques. L'étude quantitative et qualitative de la mise en visibilité et de la valorisation de ces données permet d'approcher les stratégies des producteurs institutionnels tandis que l'analyse du contenu des portails gouvernementaux, et en particulier des infrastructures nationales de données géographiques, met en évidence de nouveaux enjeux relatifs à la gestion des flux d'information.

ABSTRACT. Institutional base maps are today compete with a variety of datasets from the Internet. Faced with new actors (multinational corporations such as Google to communities of practice such as OpenStreetMap), historical spatial data producers need to reaffirm the legitimacy of their data. Our proposal is to analyze their strategies to (re)positioning on the web. From a corpus of 1674 sites providing environmental open data in Brazil, Bolivia and Argentina, we analyze the particular case of spatial data. Quantitative and qualitative study of visibility and value of these data allows us to approach the strategies of institutional producers. Then, the content analysis of the government portals, particularly the national spatial data infrastructures, highlights the critical importance of managing spatial data flow..

MOTS-CLES : donnée ouverte ; donnée géographique ; web ; infrastructure de données géographiques ; Argentine ; Brésil ; Bolivie.

KEYWORDS: open data ; spatial data ; web ; spatial data infrastructure ; Argentina, Brazil, Bolivia.

DOI:10.3199/JESA.45.1-n © Lavoisier 2012 [AR](#) [DOI](#)

1. Introduction : des données d'autorité contestées qui nécessitent un (re)positionnement des acteurs publics sur le web.

L'information géographique est au cœur des évolutions technologiques issues du web et de leurs usages depuis la fin des années 1990. Le développement et la diffusion des technologies géomatiques concernent désormais aussi bien les praticiens territoriaux que des acteurs individuels ou collectifs dans le cadre de démarches qui peuvent être de nature commerciale, communautaire, citoyenne, scientifique... Ce faisant, les institutions qui avaient jusqu'alors conçu et fourni les données de référence aux échelles supra-locales sont désormais contournées voire concurrencées par des acteurs aux statuts, natures et motivations très diverses : des firmes multinationales comme Google aux communautés de pratique comme OpenStreetMap (Feyt et Noucher, 2014). Dès lors, comment les autorités publiques se « (re)positionnent-elles » sur le web pour affirmer la légitimité des données d'autorité (des producteurs publiques) sur les données d'usage (des utilisateurs) ?

Pour tenter de répondre à cette question, nous formulons deux hypothèses. D'une part, la capacité à *rendre visible* leurs données à travers l'abondance de l'offre, est désormais l'un des paramètres qui fonde l'autorité des producteurs institutionnels d'information géographique. D'autre part, la *gestion des flux* (et non simplement le contrôle de la production) d'information géographique est aujourd'hui au cœur des politiques d'information des autorités publiques.

Pour tester ces hypothèses nous proposons une double analyse. Tout d'abord, nous mobiliserons un corpus de données recueilli dans le cadre du projet BAGUALA¹. Ce dernier recense et qualifie les sites Internet qui diffusent des données environnementales en libre accès dans trois pays d'Amérique Latine. Nous étudierons alors la place qu'occupent les données géographiques (2) pour évaluer leur visibilité sur le web au milieu de l'ensemble des sites environnementaux diffusant des données ouvertes (open data). Puis, nous étudierons le cas particulier des portails institutionnels sur le web qui, via la mise en place d'infrastructure de données géographiques, se (re)placent au centre des flux d'information. Nous analyserons alors le contenu des géocatalogues de trois infrastructures de données géographiques (3) pour tenter d'approcher les nouvelles stratégies de contrôle des autorités publiques. Enfin, un retour sur les hypothèses initiales permettra de mettre en exergue le potentiel heuristique de l'analyse de la place des données géographiques dans les sites web en général et des IDG en particulier, comme méthode d'identification de pratiques émergentes de politiques d'information par les acteurs publics. Plus particulièrement, elle permet une meilleure compréhension de ce que la transition numérique impose comme contraintes à l'État et ouvre ainsi de nouvelles perspectives de recherche (4).

¹ Le projet de recherche BAGUALA (2010-2014) financé par la chaire mixte « environnement et développement » (CNRS, Université Paris 1 Panthéon Sorbonne) vise à analyser les impacts de la diffusion des données environnementales en libre accès sur les pratiques de gestion de l'environnement. Pour plus d'information : <http://baguala.hypotheses.org>

2. Analyse de la *visibilité* et de la *valorisation* des données géographiques sur le web. Le cas du Brésil, de la Bolivie et de l'Argentine.

L'objectif de ce premier axe d'analyse est d'étudier la *place* des données géographiques institutionnelles sur le web pour tester la première hypothèse : l'un des paramètres qui fonde aujourd'hui l'autorité et la légitimité des producteurs institutionnels d'information géographique est lié à leur capacité à valoriser et à rendre visible leurs données en étant présent sur le web. Pour ce faire, nous mobilisons un corpus de 1674 sites web d'Amérique Latine (2.1) et développons une analyse quantitative pour évaluer la *visibilité* des données géographiques dans l'offre en données environnementales en libre accès (2.2). Celle-ci est complétée par une analyse qualitative permettant d'évaluer les spécificités de *valorisation* des données géographiques par rapport aux autres données environnementales diffusées (2.3).

2.1. Le corpus de données : 1674 sites web indexés

2.1.1. La constitution du corpus de données.

Les analyses développées ici s'appuient sur la constitution d'un corpus représentatif de l'offre de sites environnementaux argentins, boliviens et brésiliens. Celui-ci constitue un instantané du web à la date de sa constitution, entre janvier et mars 2012. A partir de requêtes-type adressées au moteur de recherche Google, ont été sélectionnés parmi les cinquante premières réponses obtenues² les sites offrant des données relatives aux aspects biophysiques de l'environnement, ou dont les auteurs proposaient des rubriques dédiées à l'environnement clairement identifiées comme telles³. Pour qu'un site soit inclus dans le corpus, il était donc nécessaire que soit clairement identifiée une volonté de la part de son auteur de rassembler des données portant sur ce que lui considérait relever de l'environnement⁴. Il devait aussi fournir au moins partiellement des données ou informations sur l'un des trois pays du corpus. Ont donc été intégrés à la fois des sites non spécialisés en environnement, mais dont une rubrique portait sur un thème environnemental, et des sites fournissant des données sur l'un des trois pays, mais dont les auteurs leur étaient étrangers, aboutissant à un nombre total de 1674 sites. Un postulat essentiel aux analyses qui suivent est que l'inventaire réalisé est représentatif des principaux

² Malgré les limites méthodologiques inhérentes à son utilisation (notamment la personnalisation algorithmique des résultats qu'il intègre), le moteur de recherche Google a été choisi car il est le plus utilisé (65% de parts de marché d'après l'étude ComScore QSearch parue en décembre 2012). La sélection des 50 premiers résultats (hors liens sponsorisés) obtenus permet d'inventorier les sites Web ayant le plus d'autorité selon l'algorithme PageRank de Google. Celui-ci s'appuie sur une conception de la qualité des documents, inspirée par les métriques d'autorité issues de la communauté scientifique (classement selon le nombre de citations).

³ Les requêtes ont associé le nom d'une entité géographique à un mot-clé thématique (biodiversité, climat, eau, déchets, conflit environnemental, etc.) et à une extension (l'une des quatre suivantes : .org, .com, .gouv, .blog). La méthode est détaillée dans (Gautreau et al., 2013).

⁴ Cette volonté est détectée à partir de la page initiale du site où sont détaillés les objectifs du site et de leurs auteurs et/ou dans la présence de rubriques dédiées au sein du site dans le cas des sites non spécialisés en environnement, mais qui en traitent dans une partie de leurs pages : dans ce second cas, c'est la présence d'un principe de classification des contenus du site, et l'identification d'une catégorie relevant de l'environnement par les auteurs eux-mêmes, qui permet d'inclure le site dans le corpus.

traits du « web environnemental » des pays concernés, c'est-à-dire de l'ensemble des sites qui parlent et débattent d'objets qui sont communément acceptés dans le débat public comme relevant de l'environnement aujourd'hui : les dimensions biophysiques de l'espace terrestre, leur dynamique, leur caractère de ressource ou de risque pour la société. Le corpus formé permet donc de caractériser à grands traits l'espace public environnemental virtuel des pays étudiés (Rogers, 2010).

2.1.2. *Le traitement du corpus : qualification et formalisation du graphe des sites*

Le premier traitement du corpus a consisté à catégoriser chaque site, à partir d'une grille de variables permettant de caractériser simultanément les sujets abordés, l'auteur du site, et de qualifier les dispositifs techniques informationnels (format des données téléchargeables et modes d'interaction avec l'utilisateur). Dans un second temps, le réseau formé par les hyperliens reliant entre eux les sites du corpus (graphe) a été déterminé par crawling⁵, et formalisé grâce au logiciel Gephi (figure 1). Une partie des sites du corpus n'est pas connectée au graphe (sites isolés), qui ne réunit que 1184 sites. La formalisation de ce graphe permet d'une part de repérer les grandes logiques de structuration des webs des trois pays, et notamment de comparer la place des différents acteurs dans leur organisation (Gautreau, à paraître). D'autre part, elle permet la mesure de « l'autorité » relative des sites, à partir de la comptabilisation des « liens entrants », c'est-à-dire du nombre d'hyperliens pointant vers chaque site.

Cette analyse se fonde sur l'idée aujourd'hui communément acceptée que les réseaux de sites Internet représentent des réseaux d'affinités entre acteurs, et peuvent être considérés comme une bonne approximation à l'identification de réseaux d'échange d'information et de relations entre groupes sociaux, le fait de créer un lien sur son site vers un autre site ayant une signification sociale forte et non due au hasard (Kleinberg et Lawrence 2001 ; Plantin, 2013). L'utilisation du « degré entrant » comme indicateur de la reconnaissance dont jouit un site et ses auteurs dans un réseau est maintenant courante pour détecter différents niveaux d'autorité ou légitimité dont il jouit à propos d'un thème donné (Adamic et Glance 2005, Gibson et al., 1998).

L'objectif de cet article n'est pas de commenter l'ensemble des résultats ou de les comparer par pays⁶ mais bien de focaliser l'analyse sur le cas des données géographiques. Ainsi, nous utiliserons ce corpus pour explorer la place des données géographiques dans l'offre en données en libre accès (2.2) et pour analyser les liens potentiels entre autorité d'un site et présence de données géographiques (2.3).

⁵ Le crawling consiste à détecter les hyperliens reliant des sites entre eux par exploration automatique du web. Dans notre cas, le crawling a été réalisé grâce au logiciel *issuercrawler*, développé par la Fondation Govcom.org (Amsterdam), dirigée par Richard Rogers. L'algorithme utilisé a recherché les liens existants entre les sites de notre corpus (logiciel *IssueCrawler*, appliqué en juin 2012, algorithme inter-actor, profondeur 2). En considérant le type de site web inclus dans l'échantillon, une profondeur de 2 a été suffisante pour accéder à l'essentiel du contenu des sites, en réduisant le temps de prospection.

⁶ L'analyse détaillée, incluant une comparaison des résultats par pays, a déjà fait l'objet d'une publication récente (Gautreau et al., 2013)

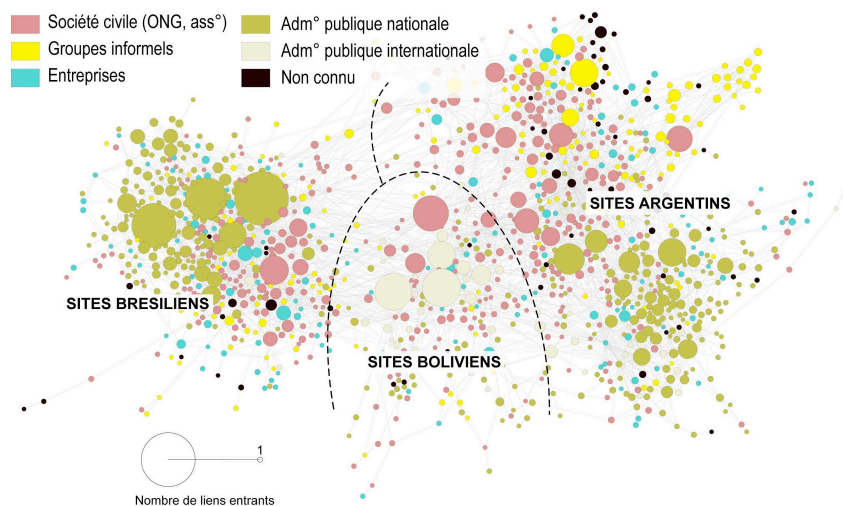


Figure 1. Graphe des webs environnementaux argentin, bolivien et brésilien
(Gautreau et al., 2013)

2.2. Analyse quantitative : quelle place occupent les données géographiques dans l'offre en données en libre accès ?

Plusieurs critères permettent de qualifier les 1674 sites diffusant de l'information environnementale au Brésil, en Argentine et en Bolivie. Parmi ces critères figure la variable *format* qui a été déclinée en huit modalités, permettant de distinguer les données en téléchargement. Sans rentrer dans les questions d'interopérabilité sémantique ou de licence juridique, il s'agit, dans un premier temps, d'analyser les niveaux de structuration et de technicité des données en différenciant les formats « conventionnels » (textes, images, cartes statiques) des formats « techniques » (vidéo, audio, cartes dynamiques). Enfin, les formats « réutilisables » désignent des ensembles de données pouvant, une fois téléchargés par l'utilisateur, être intégrés dans ses propres bases de données ou faire l'objet de combinaison numérique avec d'autres ensembles de données aux formats compatibles : il s'agit pour l'essentiel des jeux de données statistiques ou géographiques.

2.2.1. La faible proportion de données géographiques « réutilisables »

Les débats actuels autour de l'*open data*, ont tendance à se focaliser sur la mise en libre accès de données dites « réutilisables », c'est-à-dire dont la structuration et le format offrent des conditions nécessaires et suffisantes à leur exploitation dans de nouveaux contextes techniques et pour des usages variés. Force est cependant de constater leur très faible part dans les sites web étudiés (figure 2). Sur les 1674 sites analysés, 38% diffusent des données statistiques et/ou géographiques. Mais parmi

ceux-ci, seul un tiers diffuse ces données sous la forme de fichiers réutilisables, soit à peine 12% du corpus global. Les deux tiers restants sont diffusés sous la forme d'applications encapsulées qui permettent la consultation statique⁷ ou interactive⁸ sans offrir de possibilité de téléchargement des données. Parmi les sites qui diffusent des fichiers réutilisables, 50% d'entre eux offrent l'accès à des fichiers statistiques⁹, 40% rendent accessibles des données géographiques¹⁰, tandis que 10% des sites proposent un téléchargement de ces deux types de données.

On ne peut cependant en conclure que les sites web environnementaux sud-américains soient dépourvus de données. Les fréquences des données conventionnelles (texte et audiovisuel) montrent que les sites ne présentant aucune donnée en téléchargement sont minoritaires (de l'ordre de 20 à 30%) : la communication au travers de sites environnementaux s'appuie clairement sur la mise à disposition de livres ou rapports numériques, ou d'images et de vidéos.

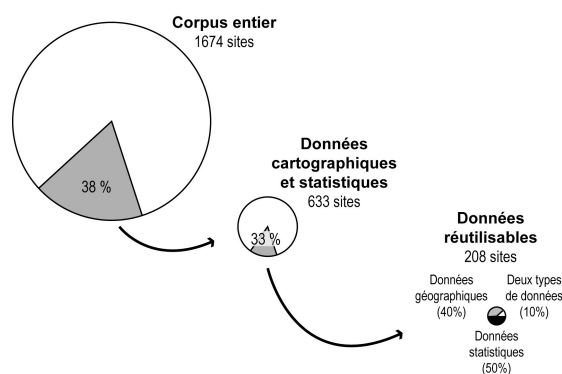


Figure 2. Disponibilité de données (géographiques et statistiques) réutilisables.

2.2.2. Une corrélation entre institutions publiques et données géographiques

L'analyse de la fréquence de disponibilité de différents formats de données dans les sites met en évidence la singularité des données géographiques et statistiques réutilisables (figure 3). Ainsi, assez logiquement, on peut observer que cette fréquence décroît proportionnellement à son niveau de technicité, et donc au niveau

7 Voir la cartothèque du Centre de Documentation et d'Information de la Bolivie (<http://www.cedib.org/mapas>) qui diffuse près de 90 cartes sous la forme fichiers JPG ou le CEDLA qui diffuse des tableaux de bord statistiques sur la politique énergétique de la Bolivie sous la forme de rapports Pdf (<http://plataformaenergetica.org>).

8 Comme l'Observatoire des Conflits Miniers d'Amérique Latine (<http://conflictosmineros.net/>) qui propose une cartographie dynamique avec frise chronologique des conflits (<http://ocmal.ourproject.org/>).

9 Cas du Centre de prévision et d'étude climatique brésilien qui permet la consultation et le téléchargement des observations issues des stations météo depuis le site : <http://bancodedados.cptec.inpe.br/>.

10 Cas de l'infrastructure de données géographiques brésilienne – INDE – qui dispose d'un catalogue permettant la consultation des métadonnées et le téléchargement de certains jeux de données <http://inde.gov.br>.

de compétences nécessaires à son utilisation (par l'auteur du site ou par ses usagers). Alors que les formats textuels ou audiovisuels sont présents dans environ 60% des sites, les cartes statiques n'apparaissent que dans 30% de ceux-ci, les jeux de données statistiques dans 10% des cas, les jeux de données géographiques dans 6%.

Ces résultats contredisent en partie l'idée communément admise que l'on assiste à un développement puissant de mise en ligne de cartes sur Internet, notamment via l'universalisation supposée du recours à l'API Google Maps. Ces analyses, il est vrai, ont jusqu'ici surtout concerné l'Europe et l'Amérique du nord. Globalement, les sites de l'administration publique présentent plus fréquemment des données à télécharger, quel que soit le format (à l'exception notable des formats audio et vidéo). Mais plus la technicité du format augmente, plus la différence avec les autres sites (gérés par des auteurs privés ou associatifs) s'accroît, témoignant du poids des capacités économiques et organisationnelles dans la mise à disposition de ce type de données, dont la production et la maintenance en ligne sont coûteux. La fréquence de données cartographiques statiques est ainsi de moitié supérieure dans les sites publics par rapport aux autres, et de quatre à six fois supérieure lorsqu'il s'agit de données statistiques ou géographiques.

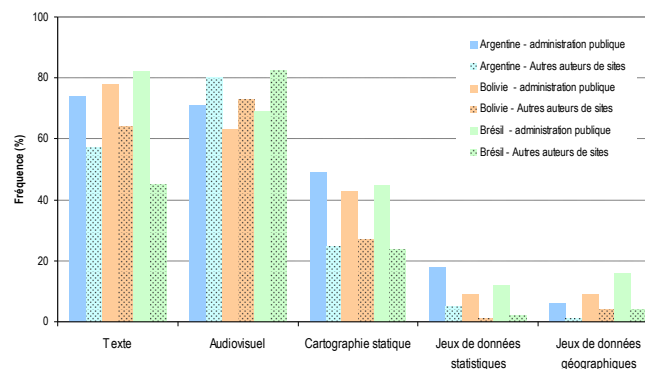


Figure 3. Fréquence de données en téléchargement selon l'auteur du site et le pays

2.3. Analyse qualitative : comment les données géographiques sont valorisées ?

C'est au niveau du site lui-même que peut être menée une analyse plus fine des fonctions attribuées aux données dans les stratégies de communication. Les pratiques éditoriales sont ainsi évaluées à partir d'une approche qualitative (2.3.1) appliquée à un échantillon de sites diffusant des données géographiques (2.3.2).

2.3.1. Pratiques éditoriales des sites autour de la « donnée »

Afin de mieux comprendre le rôle attribué aux données dans les sites environnementaux sud-américains, il est nécessaire de changer de niveau d'observation et de décrire les stratégies informationnelles à l'échelle du site. Pour

ce faire, les pratiques éditoriales de trois catégories différentes de sites ont été analysées : ceux spécialisés dans les questions de conservation, ceux fournissant des données géographiques, et ceux dont l'objectif est l'activisme social (revendication de droits et campagnes de protestations relatives à des injustices environnementales). Ces trois catégories présentent l'intérêt d'offrir une diversité d'acteurs (des sites institutionnels aux réseaux informels) et de contenus (des données expertes au contenu grand public). Pour chaque catégorie, un échantillon de 30 sites possédant la plus forte autorité dans le corpus a été exploré et décrit selon une grille à deux dimensions. L'analyse du contenu éditorial et de l'organisation du site (visibilité) permet d'évaluer le poids de la donnée dans le discours et la volonté de mettre en avant sa diffusion. L'analyse de la forme et de l'ergonomie des interfaces d'accès aux données (valorisation) permet d'évaluer les modalités opérationnelles cherchant à améliorer la valeur de la donnée. Cette démarche consiste à repérer ce que Weltevrede (2009) nomme les « arrangements techniques » utilisés par les acteurs des webs environnementaux, et à proposer une interprétation de la fonction stratégique (et politique) que leurs auteurs leur attribuent. Nous relatons dans la partie suivante les résultats propres à l'échantillon des 30 sites ayant le plus d'autorité qui diffusent des données géographiques réutilisables.

2.3.2. *Formes de valorisation de la donnée géographique*

La consultation des sites diffusant des données géographiques montre que celles-ci font l'objet d'une forte mise en visibilité. Les activités de production et diffusion de données géographiques sont évoquées dans les pages de présentation du site qu'il s'agisse ou non du cœur de métier de l'organisation. Les données géographiques sont rarement dispersées : des rubriques leur sont dédiées, accessibles par liens rapides (onglet « mapas » de plusieurs sites brésiliens). Sur tous les sites étudiés, l'accès aux données s'effectue rapidement, en très peu de clics. Cette mise en visibilité de données plutôt techniques qui ne s'adressent, par conséquent, qu'à un public limité, témoigne de la volonté de mise en exergue des investissements et des compétences techniques des organismes diffuseurs. Même quand les jeux de données diffusés sont relativement limités (quelques couches SIG superposées à des référentiels nationaux) la publicité autour des données géographiques est importante et semble vouloir renforcer la légitimité d'action des diffuseurs.

La donnée géographique n'est quasiment jamais diffusée sans une mise en valeur de son contenu. Cette dernière passe par une mise en cartes qui permet de combiner le jeu de données diffusé avec d'autres sources ou de le mobiliser pour réaliser des analyses thématiques. 25 des sites analysés encapsulent la donnée dans des interfaces statique (cartothèque) ou dynamique (WebMapping, API Google Maps). Mais, seuls les 2/3 des sites proposent des métadonnées (des fichiers PDF non normés aux XML intégrés dans un géocatalogue). Une part non négligeable de données est donc diffusée sans aucune documentation. Pourtant, la mise à disposition des données semble, au niveau du contenu éditorial et de l'arborescence du site, tout aussi centrale. Enfin, la moitié des données géographiques diffusées est accompagnée de commentaires, la plupart du temps sous la forme de publications.

L'exploration de l'échantillon met en évidence deux types de stratégies autour de la donnée géographique (figure 4). Celle, d'abord, des sites qui diffusent des données faiblement valorisées (non documentées, non analysées) mais qui disposent d'une forte visibilité. Ce type regroupe 6 des 30 sites analysés qui semblent faire de la donnée un « *produit d'appel* » c'est-à-dire un support visant à attirer les visiteurs en s'affichant en *vitrine* (dans les pages principales) du site. La seconde stratégie correspond aux sites qui diffusent des données géographiques ayant une bonne visibilité dans l'arborescence du site et faisant l'objet d'un effort de valorisation *via* la mise en ligne d'un géocatalogue, et la *mise en scène* des données dans un système d'information en ligne (sous-domaine spécifique). Ce second type regroupe 21 des 30 sites de l'échantillon. Il concerne à la fois des dispositifs de type observatoire dont l'analyse est le cœur de métier et des Infrastructures de Données Géographiques dont le partage de données est le cœur de métier.

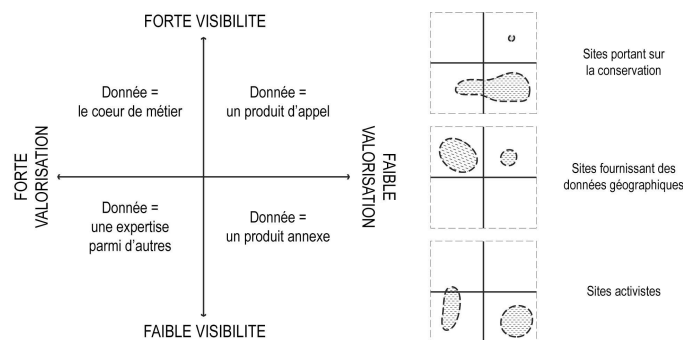


Figure 4. Analyse de la place des données sur trois types de sites étudiés.

2.4. Conclusion intermédiaire.

L'analyse d'un ensemble de sites web diffusant des données environnementales permet de prendre du recul par rapport au cas particulier des données géographiques. Il révèle le (re)positionnement des institutions et leurs stratégies de mise en visibilité pour conforter (ou renforcer) la légitimité de leur production. Ces analyses mettent en évidence trois caractéristiques permettant de mieux appréhender les nouveaux régimes de diffusion de l'information géographique (Joliveau et al., 2013).

Tout d'abord, si les données géographiques (ou statistiques) réutilisables occupent une place centrale dans les débats autour de l'*open data*, elles restent très largement minoritaires dans les sites web environnementaux sud-américains. Il existe donc un volume conséquent de données non normées et difficilement réutilisables avec les outils actuels mais qui constitue pourtant une ressource potentiellement riche (et hétérogène) de connaissances sur l'environnement. Par ailleurs, c'est fréquemment sur des sites d'administration publique que ces données sont diffusées. Ces derniers disposent d'une forte autorité sur le réseau. Ainsi, bien

que minoritaires les données géographiques occupent une place non négligeable sur Internet. Enfin, la mise en ligne de ces données par les autorités publiques est caractérisée par une mise en scène et une valorisation conséquente via l'utilisation de géocatalogues, de géoportails, de cartothèques ou d'API. Le développement d'Infrastructure de Données Géographiques (IDG) en est un bon témoin.

3. Analyse des modes de gestion des flux de données géographiques. Le cas de trois infrastructures nationales de données géographiques d'Amérique Latine.

3.1. Analyser les IDG pour révéler les stratégies de diffusion des données géographiques institutionnelles

Le besoin exprimé au niveau international de faciliter l'accès, l'utilisation et le partage des données géographiques a conduit depuis les années 1990 au développement d'IDG nationales. Ces plateformes rassemblent les données, les réseaux informatiques, les normes et standards, les accords organisationnels et les ressources humaines nécessaires pour faciliter et coordonner le partage, l'accès et la gestion des données géographiques (Rajabifard *et al.*, 2002). D'un point de vue technique, les IDG peuvent être assimilées à des systèmes d'information bâtis autour d'une architecture orientée services dont la composante web est essentielle. Leur mise en réseau, du fait de l'interopérabilité et de la normalisation des données, facilite la circulation de contenus grâce, par exemple, au « moissonnage » de catalogues émanant de producteurs indépendants.

Dès lors, les IDG apparaissent comme des objets pertinents pour avancer sur les questions liées aux nouvelles formes de diffusion de l'information géographique par les autorités publiques (Noucher, 2013). Nous proposons ici de les mobiliser afin d'évaluer les patrimoines de données géographiques disponibles et d'appréhender alors les stratégies informationnelles qu'elles traduisent. A partir du décryptage de leur géocatalogue (3.2.), l'analyse de l'emprise spatiale des données (3.3) tente d'initier une géographie de l'information géographique institutionnelle (3.4).

3.2. La constitution du corpus de données.

Une chaîne de traitement pour l'extraction automatisée des métadonnées et leur restructuration dans une base de données a été développée pour l'analyse spatio-temporelle des infrastructures de données géographiques (Pierson *et al.*, 2013). L'extraction des fichiers XML de métadonnées est opérée à partir de scripts en Python. Des requêtes XQuery sur les 90 balises XML identifiées comme similaires aux trois catalogues permettent ensuite de structurer les données dans une base PostGIS. L'analyse (statistique ou cartographique) est réalisée à partir de scripts en Python ou de traitement dans QGIS. Cette chaîne a été appliquée aux Infrastructures Nationales de Données Géographiques (INDG) des trois pays étudiés (figure 5).

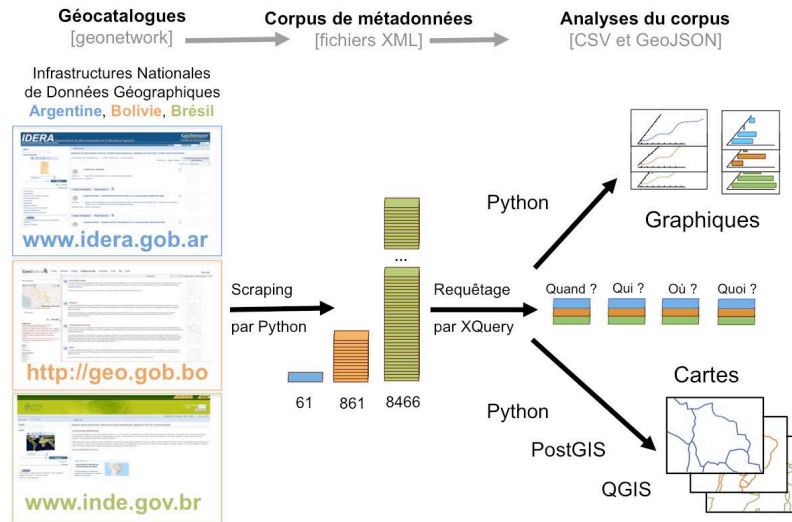


Figure 5. Chaîne de traitement pour l'analyse des trois INDG

On remarque d'emblée une très forte disparité dans le volume de ces catalogues qui s'explique par des historiques différents : au Brésil l'*Infraestrutura Nacional de Dados Geoespaciais* est ouverte depuis de 2007 ; en Bolivie, le catalogue de *Geobolivia* est le fruit d'un travail d'inventaire et de documentation systématique entrepris depuis 2012 par la Vice-Présidence (Lerch, 2013) ; enfin en Argentine, l'*Infraestructura de Datos Espaciales de la Republica Argentina*, pilotée par l'Institut Géographique National argentin, n'a ouvert qu'en septembre 2013. La comparaison d'INDG ayant des niveaux de maturité différents semble intéressante pour tenter d'identifier les trajectoires de montée en charge de ces dispositifs institutionnels. L'interrogation des champs de métadonnées (balises des fichiers XML) permet de traiter en particulier de la couverture temporelle (quand ?), organisationnelle (qui ?), thématique (quoi ?) et géographique (où ?) des données. Seule cette dernière dimension sera illustrée à la section suivante.

3.3. Analyse de l'emprise spatiale des données diffusées.

L'extraction des 4 coordonnées des rectangles d'emprise (un même jeu de données peut posséder plusieurs emprises si la couverture est discontinue) est réalisée à partir de la récupération des balises XML correspondantes aux emprises (`<EX_GeographicBoundingBox>`) en format CSV pour générer ensuite des données en format GeoJSON. Un traitement dans PostGIS/QGIS, permet enfin, pour chaque pays, de produire une carte des densités de données géographiques. Cette « géographie de l'information géographique » révèle alors des répartitions hétérogènes de l'information et des discontinuités spatiales qui ne sont pas le seul fait de disparité démographique (figure 6).

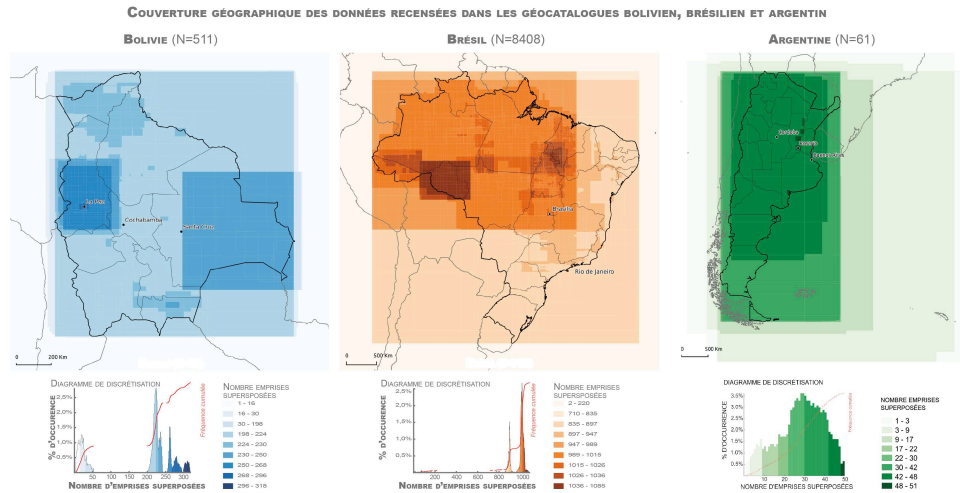


Figure 6. Couverture spatiale des données diffusées dans les trois INDG étudiées (Extraction en avril 2013 - Bolivie et Brésil - et en octobre 2013 - Argentine)

Ainsi, en Bolivie, l'aire urbaine de Cochabamba (la 2^{ème} du pays) apparaît comme beaucoup moins couverte que des secteurs quasi désertiques comme l'Oruro ou le Potosi. Au Brésil, le massif amazonien est plus couvert que la bande littorale. L'intensité historique des recherches sur l'environnement amazonien et le grand nombre de projets de développement et d'aménagement du territoire mis en place dans cette région ne sont sans doute pas étrangers à cette différenciation de la couverture géographique des données. Enfin, en Argentine la couverture en apparence plus homogène du territoire par l'IDERA cache le fait qu'il s'agit d'une INDG très récente, où n'ont été incorporés jusqu'à maintenant qu'un nombre réduit de jeux de données, pour l'essentiel de portée nationale. L'analyse de la couverture géographique des données cartographiques institutionnelles permet donc de révéler des situations potentiellement intéressantes à approfondir pour comprendre les dynamiques territoriales et/ou politiques de constitution de ces patrimoines.

3.4. Conclusion intermédiaire

L'analyse du contenu des INDG du Brésil, de Bolivie et d'Argentine révèle que ces plateformes ne se contentent pas de diffuser des référentiels géographiques nationaux (à l'exception de l'Argentine). Les données proposées ne couvrent pas de manière homogène et égalitaire l'ensemble du territoire. Si elles donnent accès à des jeux de données nationaux, elles offrent aussi un lien vers des données produites plus ponctuellement. On note d'ailleurs que plus l'INDG est ancienne plus la distribution spatiale des données diffusées se fait selon des répartitions aux deux extrêmes, se concentrant donc sur des données nationales et locales (figure 7). Ainsi, les institutions qui jadis affirmaient leur légitimité par leur capacité à produire de

manière homogène des référentiels sur l'ensemble du territoire modifient aujourd'hui, sur le web, leurs arguments d'autorité ; On peut faire l'hypothèse que ces initiatives cherchent désormais à mettre en avant leur capacité à gérer les flux d'information géographique et à devenir de véritables « hubs » dans un réseau qui se complexifie, s'intéressant aussi bien à des données qui couvrent l'ensemble de leur territoire qu'aux données locales qui le morcellent.

Distribution du taux de couverture spatiale des données géographiques diffusées par trois Infrastructures Nationales de Données Géographiques (Argentine, Bolivie, Brésil)

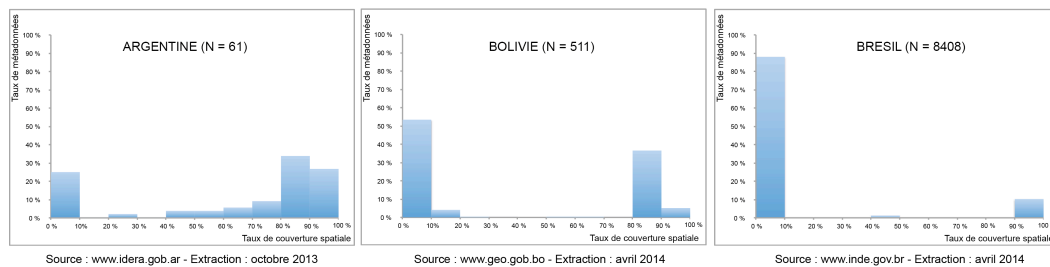


Figure 7. Distribution du taux de couverture spatiale dans les trois INDG étudiées¹¹.

4. Conclusion : le web comme (un des) terrains

L'objectif de ces analyses était de saisir sur le terrain les manières par lesquelles les données géographiques institutionnelles se (re)positionnent sur Internet pour faire autorité. Le terrain est ici à comprendre comme l'exploration du web à partir d'analyses quantitatives (comparaison du volume de données géographiques par rapport au volume de données ouvertes tous formats confondus), d'observations qualitatives (qualification des pratiques éditoriales autour de la donnée géographique) ou d'extraction et de traitement de données (*web scraping* des géocatalogues pour cartographier les métadonnées).

Ce faisant, les résultats obtenus nous semblent révéler des éléments intéressants sur la place des données géographiques dans la gouvernance informationnelle de l'environnement (Gautreau et Noucher, 2013) et en particulier sur les stratégies des autorités publiques pour rester au centre des flux. Mais le web comme terrain nécessite désormais d'être complété par une analyse « incarnée » des usages de ces données en libre accès. Il s'agit ainsi de prolonger l'analyse exploratoire du web environnemental et des IDG par une enquête auprès des utilisateurs pour comprendre via l'analyse des modalités de remobilisation des données mises à disposition, les effets de ces recompositions informationnelles. Pour ce faire, les

¹¹ Pour chaque jeu de données, on calcule la part du territoire national qu'il recouvre. Un jeu de donnée couvrant moins de 10% du territoire national correspond généralement à une information locale. Inversement, un jeu de données couvrant plus de 80% du territoire correspond à une information de portée nationale. On peut alors, en calculant la distribution des jeux de données en fonction de ce paramètre, caractériser les patrimoines de données des IDG en fonction des niveaux géographique de couverture de leurs données.

résultats présentés ici apparaissent comme un bon matériau de recherche permettant de poser de nouvelles hypothèses cherchant, cette fois, à faire parler les usages. Ils ouvrent ainsi des perspectives de recherche complémentaires.

Bibliographie

- Adamic L. et Glance N. (2005). « The Political Blogosphere and the 2004 U.S. Election: Divided They Blog », *3rd international workshop on Link discovery*, pp.36-43
- Feyt G. et Noucher M. (2014). La gouvernance informationnelle, outil et enjeu stratégiques des recompositions territoriales : vers l'émergence de nouveaux référentiels géographiques ? *Colloque CIST "Fronts et frontières des sciences du territoire"*, Paris, pp. 191-196.
- Gautreau P. (à paraître). Etat, information environnementale et pouvoir : ce que change Internet en Argentine, Bolivie et Brésil. *Actes du Colloque d'Orléans « Géographie, écologie, politique : un climat de changement »*, 6-7 octobre 2012.
- Gautreau P., Severo M., Giraud T. et Noucher M. (2013). « Formes et fonctions de la 'donnée' dans les webs environnementaux sud-américains (Argentine, Bolivie, Brésil) » *Networks and communication studies*, vol. 27/1-2, pp. 22-59.
- Gautreau P. et Noucher M. (2013), « Gouvernance informationnelle de l'environnement et partage en ligne des données publiques. », *Networks and communication studies*, vol. 27/1-2, pp. 5-21.
- Gibson D., Kleinberg J. et Raghavan P. (1998), « Inferring Web communities from link topology », *Proc. 9th ACM Conference on Hypertext and Hypermedia*, 20-24 juin 1998, Pittsburgh, USA.
- Kleinberg J. et Lawrence S. (2001). « The Structure of the Web », *Science*, n°294, p. 1849.
- Lerch L. (2013). « Logique de projet et régulation publique de l'information géographique : l'expérience bolivienne », *Networks and communication studies*, vol. 27/1-2, pp. 99-119.
- Noucher M. (2013), « Infrastructure de données géographiques et flux d'information environnementales : de l'outil à l'objet de recherche », *NETCOM Networks and communication studies*, vol. 27, n°1-2, pp. 120-147.
- Pierson J., Noucher M., Gautreau P., Lerch L., Pissot O., Jautard A. et Lesage S. (2013). « Analyse des patrimoines nationaux de données géographiques. Comparaison de trois infrastructures de données géographiques (Bolivie, Brésil, France) », SAGEO'13, Brest.
- Plantin J-C (2013). « D'une carte à l'autre : le potentiel heuristique de la comparaison entre graphe du web et carte géographique », in *Analyser le Web en sciences humaines et sociales*, Barats C. (ed.), Armand Colin, pp. 228-242
- Rajabifard A, Feeney M.-E.F., Williamson I. et Masser I. (2002), « Chapter 6, National SDI Initiatives, in Williamson I., Rajabifard A. et Feeney M.-E.F. (eds), Development of Spatial Data Infrastructures: from Concept to Reality, Taylor & Francis, pp. 95-109.
- Rogers R. (2010), "Mapping public Web space with the Issuecrawler", in: Brossard C., Reber, B. (Eds.), Digital Cognitive Technologies: Epistemology and Knowledge Society, Wiley, Londres, pp.115-126.
- Weltevrede E. (2009) Thinking Nationally with the Web. A Medium-specific Approach to the National Turn in Web Archiving. University of Amsterdam, Amsterdam.